# Variational Encoder Decoder for Image Generation conditioned on Captions

Nicolas Angelard-Gontier, Joshua Romoff, Prasanna Parthasarathi

Reasoning and Learning Lab
McGill University

IFT 6269 - Probabilistic Graphical Models
Fall 2017

RLLab
McGill University

# Introduction

- This poster presents the work we did for the Probabilistic Graphical Models class.
- Inspired by Variational Auto Encoders, we decided to look at a Variational Encoder Decoder model to generate images based on captions.
- We experimented on the MNIST dataset with custom operation captions.

RLLab
McGill University

# MNIST Dataset

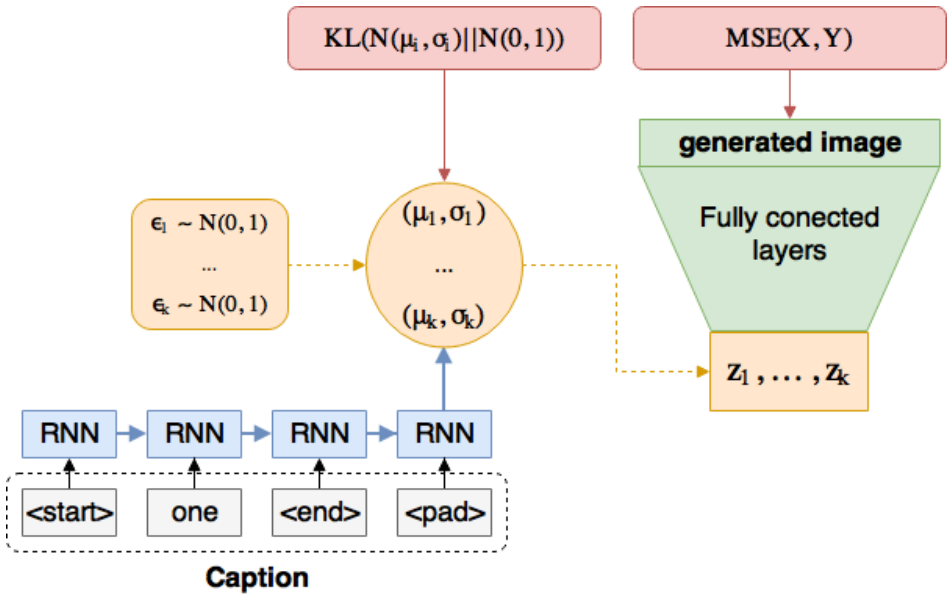
We generated digit images (1x28x28) from MNIST by giving hand-crafted captions. We manually created four types of captions for this dataset:

1. image label as caption, ie: "<start> *five* <end>" for all images.
2. random sentences with one label in it, ie: "<start> this is a black *five* on a white background <end>". We have a total of 12 different sentences, and for each image, we sample one of them to be the caption.
3. logical captions with multiple numbers in it, ie: "<start> min nine *five* seven <end>" should generate an image of a five. We considered 'min' and 'max' operators.
4. operation captions where we write a sum of multiple numbers that represent the image label, ie: "<start> three plus two <end>" should generate an image of a five.

RLLab McGill University

# Model

1. Encoding: Recurrent Neural Network is fed word embedding vectors for each word in the caption $(w_0, \ldots, w_N)$, and encodes the caption to a vector representation $(c)$ in a high dimension space.
2. Sampling: $c$ fed to a one-layer fully connected network to generate a set of $\mu_i \in \mathbb{R}^{dim}$ and $\sigma_i \in \mathbb{R}^{dim}$ vectors $\forall i \in \{1, \ldots, k\}$.
   - independent gaussians: sample $z_i = \mu_i + \epsilon_i \sigma_i$ with $\epsilon_i \sim N(0, 1) \in \mathbb{R}^{dim}$
3. Generate: $(z_1, \ldots, z_k)$ fed to a multi-layer fully connected network to generate a vector of size 27 x 27 representing a digit image.



# Training Loss

We used the Mean Squared Error loss between the generated image and the true image as our reconstruction objective

Furthermore, we regularized the encoder with the KullbackLeibler divergence loss between $N(\mu_i, \sigma_i)$ and $N(0, 1)$.

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 + KL(N(\mu_i, \sigma_i^2)||N(0, 1))$$

RLLab
McGill University

# Experiments

We trained the VED network on captions made of only one elementary operation.

<start> five plus two <end>

We tested the trained network on unseen, more complicated captions:

<start> two plus one plus two <end>

<start> one plus five plus two plus one <end>

RJLab McGill University

# Future Work

- Attention mechanism on the Encoder network.
- Smaller and predefined word embeddings.
- Experiments with the Cross Entropy loss instead of the MSE reconstruction loss.
- Deconvolution layers instead of multi-layer fully-connected layers.

RLLab
McGill University